In: D. Krallmann (Hrsg., 1978): Kolloquium zur Lage der linguistischen Datenverarbeitung. LDV-Fittings e.V. Essen

H. Zimmermann

AUTOMATISCHE TEXTANALYSE UND INDEXIERUNG
(LDV-Aktivitäten in Regensburg - 1. Teil)

0 Vorbemerkung:

Die Darstellung der LDV-Aktivitäten an der Universität Regensburg ist in zwei Teile gegliedert. Der erste Teil bringt eine Übersicht über die einzelnen Schwerpunkte und im Anschluss eine ausführlichere Information zu dem Forschungsprojekt JUDO sowie eine kurze Darstellung des Projekts ENZY. Im Teil 2 (vgl. Krause 1978) wird das Projekt COBAPH vorgestellt und v.a. auf neuere Entwicklungen von DONALD hingewiesen.

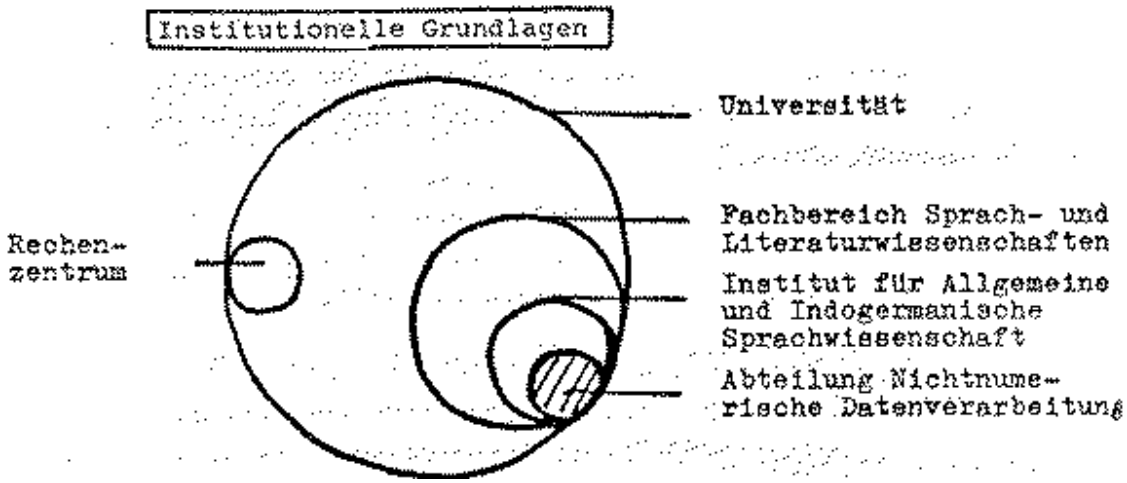
1 Übersicht zu den LDV-Aktivitäten an der Universität Regensburg:

1.1 Institutionelle Grundlagen:

Die Universität Regensburg besitzt keinen Informatik-Schwerpunkt. Sie verfügt jedoch über ein Rechenzentrum (mit Siemens 4004-45 und TR 440 Doppelprozessor) als zentrale (Service-) Einrichtung; im Rahmen der Betriebswirtschaft ist ein spezieller Lehrstuhl für Wirtschaftsinformatik eingerichtet und de facto wird auch eine Ausbildung in Rechtsinformatik durchgeführt. Diese Orientierung auf "Angewandte Informatik" wird ergänzt um eine "Abteilung für Nichtnumerische Datenverarbeitung" (NDV), der eine entsprechende "Abteilung für Numerische Datenverarbeitung" folgen sollte (bisher aber den Haushaltseinsparungen zum Opfer fiel).

Die Abteilung NDV ist dem Fachbereich "Sprach- und Literaturwissenschaften" zugeordnet und dort wiederum integriert in das "Institut für Allgemeine und Indogermanische Sprachwissenschaft". Aus dieser Orientierung versteht sich die Betonung der "Linguistischen Datenverarbeitung" in den Aktivitäten der Abteilung.

Personell ist die Abteilung ausgestattet mit einer HS-3-Stelle (Wiss. Rat und Professor Dr. H. Zimmermann) und zwei BAT IIa/Ib Stellen (Dr. J. Krause, L. Hitzenberger, M.A.), wovon die letztere derzeit aus Einsparungsgründen zur Hälfte gesperrt ist. Daneben verfügt die Abteilung über 1/2 Sekretärin und eine wiss. Hilfskraft.



1.2 Lehre/Ausbildung:

Seit 1974 ist die HS-3-Stelle besetzt. In dieser Zeit ist es gelungen, die Ausbildung, die bis dahin in loser Form stattfand, in ein fachbezogenes Studium zu integrieren: die Abteilung vertritt das Teilfach "Linguistische Informationswissenschaft" (LI) im Rahmen des Studiums der "Allgemeinen Sprachwissenschaft". Es ist damit möglich, mit einem Studienschwerpunkt LI (und einem weiteren Teilfach der Allgemeinen Sprachwissenschaft sowie einem Zweitfach, z.B. Germanistik, aber auch z.B. - mit Ausnahmegenehmigung - Mathematik) einen regulären Studienabschluss - Magister Artium oder Promotion - zu erlangen.

Das Studienfach "Linguistische Informationswissenschaft" vereinigt Aspekte der Linguistik, der LDV, der Informatik und der Informationswissenschaft (Information und Dokumentation). (Näheres zum Konzept vgl. Zimmermann, H. 1976). Im Rahmen der LI-Komponente "EDV-Ausbildung" wird versucht, auch allgemeine Aspekte (insbesondere in Bezug auf die EDV-Grundausbildung für den Fachbereich) abzudecken; in Vorbereitung ist eine engere Kooperation im Grundlagenbereich mit der Wirtschafte- und Rechtsinformatik.

1.3 Forschung und Entwicklung:

Im Rahmen der Forschungstätigkeit der Abt. NDV sind zwei Schwerpunkte zu unterscheiden:

([1)] wird ausführlich im Referat vorgestellt, [2)] wird im Referat kurz vorgestellt)

Ein Bereich (A) beschäftigt sich mit der Entwicklung von DVSystemen bzw. Modellen, bei denen linguistische Verfahren zum Einsatz kommen.

Ein zweiter Komplex (B) befasst sich mit Hilfsmitteln (im Bereich Dokumentation und Textverarbeitung), die die Forschung und Entwicklung im Bereich A unterstützen sollen.
Zum Bereich A sind folgende Projekte zu rechnen:

(1) [1] JUDO: Modellentwicklung eines Software-Systems zur computergestützten Indexierung am Beispiel der Verarbeitung juristischer Dokumente (Förderung durch BMFT, FKZ - PT 131.04, seit Juli 1977),

(2) [2] ENZY: Automatische Bearbeitung enzyklopädischer Lexika (Entwicklungsauftrag der Fa. Siemens, Bereich Nachrichtentechnik, seit Juli 1975),

(3) ABIS: Automatisiertes Büro-Informations- System. Pilotanwendung des Systems CONDOR; Förderung durch BMFT in Aussicht; Beginn April/Mai 1978 (da dieses Projekt erst in der Planung ist, kann vorläufig keine ausführlichere Information gegeben werden),

(4) [2] DONALD, Teilbereich: linguistische Analyse juristischer Texte. Zur Finanzierung/Förderung vgl. (6)

Zum Bereich B sind folgende Forschungs-und Entwicklungsprojekte zu rechnen:

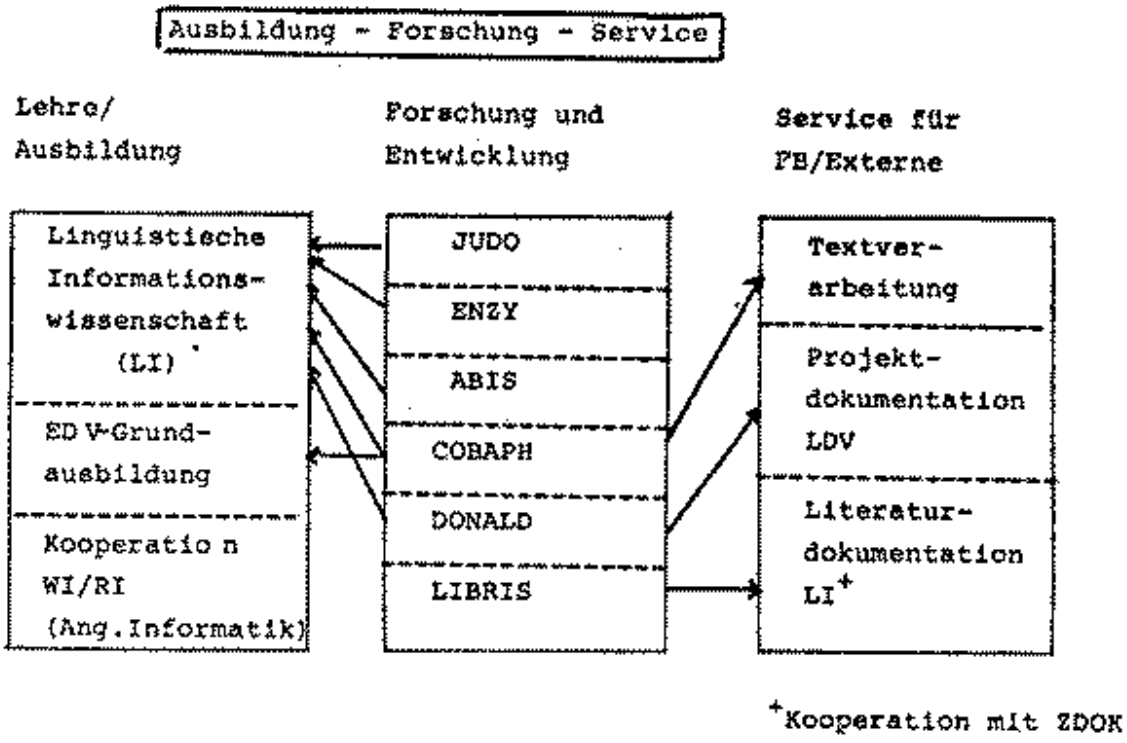
(5)[1] COBAPH: System zur computergestützten Herstellung von Textwörterbüchern, Konkordanzen ...

(6)[2] Projektdokumentation DONALD: Dokumentation von Arbeiten (Projekten) auf dem Gebiet der linguistischen Datenverarbeitung; Förderung/Finanzierung durch Fa. Siemens, D AP GE, seit 1974.

(7) LIBRIS: Literaturdatenbank zum Bereich der Linguistischen Informationswissenschaft. Verwendung des Systeme TUBIBMUE der TU München, einer speziellen TELDOK/DBS-Anwendung.
Die Datenbank soll im On-line-Retrieval solche LI-Literatur (ggf. mit Abstract/Teilüberschriften) zugänglich machen, die an der Abteilung NDV verfügbar (v.a. "graue Literatur") bzw. bei ZDOK erhältlich ist (dabei Kooperation mit ZDOK bei der Literaturerschließung).

1.4 Service-Einrichtungen:

Die Abteilung unterhält die unter 1.3 bezeichneten Hilfssysteme COBAPH, die Projektdokumentation DONALD und LIBRIS auch zur Unterstützung von Mitgliedern des Fachbereichs Sprach- und Literaturwissenschaften, der Universität und externer Interessenten. Sie berät insbesondere bei entsprechenden Vorhaben und stellt ihre Systeme zur Verfügung, ohne dass jedoch über Wartung, Benutzungsanleitung und Beratung hinaus weitere Hilfeleistungen (etwa Datenerfassung) gegeben werden können.



2 Projekt JUDO: Modellentwicklung eines Softwaresystems zur Computergestützten Indexierung am Beispiel der Verarbeitung juristischer Dokumente

Die Konzeption von JUDO wurde inzwischen mehrfach vorgestellt [1]. Daher sei hier der Versuch unternommen, einmal exemplarisch aufzuzeigen, wie sich das Verfahren anhand konkreter Texte (nach den jetzigen Vorstellungen) gestalten soll ([1] vgl. u.a. Zimmermann 1977a; Zimmermann 1977b; Krause 1977):

2.1 Eine (Fach-) Textmenge, an der wir das Verfahren entwickeln und erproben wollen, ist das EDV-Recht. Darunter fällt - als wohl wichtigste Norm (Gesetz) - das Bundes-Datenschutz-Gesetz. Es ist derzeit in 47 Paragraphen ("Dokumente") gegliedert. Wir stellen uns nun vor, dass ein JURIST (also ein Fachmann) bzw. ein JOURNALIST (also ein "halber" Fachmann) aufgrund von Begriffen (Deskriptoren), die er als repräsentativ für seine 'Problemstellung' ansieht (und die er mit den in heutigen Dokument-Retrieval-Systemen üblichen Methoden verknüpft), zu den Paragraphen (als 'Dokumenten') hingeführt wird, die ihm im Hinblick auf seine Problemstellung relevant erscheinen, ohne dass er allzu viel Ballast erhält (also unzutreffende Paragraphen) und ohne dass ihm relevante Paragraphen nicht nachgewiesen werden.

2.2 Zu diesem Zweck werden die deutschsprachigen Texte/Dokumente zunächst aufbereitet, d.h. u.a. einer maschinellen linguistischen Analyse unterzogen. Als Analyseverfahren wird die "Saarbrücker Automatische Textanalyse" (SATAN) herangezogen (die jeweils lauffähige neueste Version steht in Regensburg zur Verfügung). Da SATAN (derzeit) satzbezogen analysiert, muss - neben einigen technischen Einschränkungen wie max. Wortlänge <= 36 Zeichen der übersatzmäßige Kontext außer Betracht bleiben. Außerdem werden (bislang) einige Präkodierungen vorgenommen. So wird z.B. der Satzendepunkt eindeutig vom Abkürzungspunkt unterschieden;

eine (verblose) Überschrift erhält eine entsprechende Kennung, auch sog. Setzungen (d.h. verblose Sätze - sehr gering an Zahl) werden markiert. Dies ist natürlich nicht unproblematisch; mittelfristig ist vorgesehen, solche Restriktionen ggf. aufzuheben, doch sollen die derzeit vorgenommenen Markierungen zugleich eine gewisse Abschätzung der Effizienz und des Aufwands für die Präkodierung ermöglichen. Dies gilt auch für die Ausnutzung der Großschreibung von Nomina; hier lässt die Analyse - da beide Varianten integriert sind - eine Abschätzung des Effekts bereits zu.

2.3 Für die Zwecke der Analyse wird der Wortschatz lexikalisch unter verschiedenen Gesichtspunkten erfasst, soweit dies nicht bereits im Rahmen des Saarbrücker Projekts "Automatische Lemmatisierung" geschieht. Vorläufig sind auch solche Wörter berücksichtigt, die sich über Lexikonregeln identifizieren ließen (was übrigens in SATAN schon realisiert ist), also vor allem nominale (Augenblicks-) Komposita. Daher werden z.B. RECHTSVERORDNUNG und nicht nur RECHT und VERORDNUNG, ENTWICKLUNGSSTAND neben ENTWICKLUNG und STAND, BUNDESZENTRALREGISTERGESETZ neben REGISTER, GESETZ usf. erfasst.

Diese Wörter sind zunächst gemäß den Kodierkonventionen des "Saarbrücker Deutschen Arbeits-Wörterbuchs" (SADAW) syntaktisch zu klassifizieren, wobei jedoch zugleich weitere potentielle (also nicht nur die im Text aktualisierten) Wortvarianten, sondern solche, die prinzipiell [1)] im Text möglich erscheinen, notiert werden: kommt im Text beispielsweise das Nomen STAND vor, wird auch die Verbinformation (STAND --> STEHEN) notiert ([1)] mit Einschränkungen: veraltete oder sehr unwahrscheinliche Varianten - dies gilt besonders für den "Semantik-Teil" - bleiben vorläufig außer Betracht) .

In diesem Zusammenhang werden auch sog. inflexible "Feste Wendungen" (FW) wie: 'IM RAHMEN', 'NACH MASSGABE', 'MIT ERFOLG', 'ES SEI DENN, DASS' festgehalten, die mithilfe eines "pattern matching" bereits bei der Wörterbuchsuche identifiziert werden und die spätere Analyse entlasten sollen.

SADAW-Einträge

```
(1) HAUS        SUB NEU SG ...
(2) HAEUSER     SUB NEU PL DAT-N....      'HAUS'
(3) IM RAHMEN   PRP +GEN/+VON ...
(4) IST         FVS 3.SG PRAES            'SEIN' ...
(5) STAND       FIV 1.3.SG PRAET          'STEH' ...
(6) STAND       SUB MAS SG ...
```

2.4 Im Rahmen von SATAN werden nach dem bisherigen Stand des Verfahrens die üblichen Analyseschritte, also Homographenreduktion (DIHOM), Segmentierung (SEGMENT), Nominal- und Subsatzanalyse (NOVERA und SYNAN) durchlaufen. Damit sind - syntaktisch - Strukturbeschreibungen erfolgt, die z.B. die nominale Zuordnung von Oberflächen-Adjektiven (als Prädikate) zu den zugehörigen Nomina erkennen lassen:

(ATT) steht für 'Adjektiv-Attribut-Relation', z.B.
TECHNISCH (ATT) MASSNAHME - TECHNISCHE MASSNAHME(EN)
(GEN) steht für 'GENITIV-ATTRIBUT-RELATION', z.B.

AUSFÜHRUNG (GEN) VORSCHRIFT -AUSFÜHRUNG DER VORSCHRIFT(EN)

TECHNISCH (ATT) MASSNAHME
ORGANISATORISCH (ATT) MASSNAHME
PERSONENBEZOGEN (ATT) DATEN
(SEIN) ERFORDERLICH (ATT) MASSNAHME
ANGEMESSEN (ATT) VERHÄLTNIS

oder z.B. die Nominalattribute identifizieren:

AUSFÜHRUNG (GEN) VORSCHRIFT
VORSCHRIFT (GEN) GESETZ
ANLAGE (ZU) GESETZ
ZUSTIMMUNG (GEN) BUNDESRAT

daneben sind (in SEGMENT) bereits Verbzusätze zugeordnet, also:

BESTEHEN ... FORT - FORTBESTEHEN (vgl. § 5 BDSG)

2.5 Die zweite Phase von SATAN (Operator SEDAM) dient einmal der kontextorientierten Vereindeutigung von Begriffen, z.B.

ANLAGE (zum Gesetz) vs.
ANLAGE (DV-Gerät) vs.
....
ANLAGE (Park ...)

Nicht immer wird dies gelingen, da die kontextuelle Analyse nicht ausreicht oder nicht fein genug ist; dennoch soll mithilfe von Regeln, die im 2. heranzuziehenden Wörterbuch (SESAM) untergebracht sind, so viel wie möglich vereindeutigt werden. Insbesondere sollen auch (flektierte) mehrwortige Begriffe identifiziert werden, also z.B.

PERSONENBEZOGENE DATEN
STAND DER TECHNIK
MASSNAHMEN TREFFEN

Die "Regeln" operieren auf Wortlauten, Merkmalen und syntaktisch ermittelten Strukturen:

SESAM-Einträge

```
(1)  WENN      Prädikat JURISTISCH UND
               Subjekt PERSON ...
     DANN      Feste Wendung JURISTISCHE PERSON

(2)  WENN      Prädikat TREFF UND
               Akk.Obj. MASSNAHME (PL) ...
     DANN      Feste Wendung MASSNAHMEN TREFF
```

```
(3)  WENN      Prädikat SORG UND NICHT
               Refl.Obj. SICH ...
     DANN      Prädikat SORGE TRAG/SORG (1)
```

2.6 Dem Benutzer (= Anfragenden beim Retrieval) sollen jedoch weitere Möglichkeiten geboten werden. Dazu ist eine Strukturierung und Klassifizierung des Wortmaterials vorgesehen. Insbesondere gilt dies für:

- eine Zerlegung von Komposita in retrieval-relevante sinnvolle Segmente, z.B.:

AKTENSAMMLUNG	=>	AKTE, SAMMLUNG (§ 37)
DATENVERARBEITUNGSANLAGE	=>	DATENVERARBEITUNG, ANLAGE (i) =/ ARBEIT!
ZUGRIFFSBERECHTIGUNG	=>	ZUGRIFF, BERECHTIGUNG =/ GRIFF, RECHT!

- eine Zuordnung von Verben und (derivierten) Nomina, z.B.:

VERARBEITEN	=>	VERARBEITUNG
ÜBERPRÜFEN	=>	ÜBERPRÜFUNG
GEWÄHRLEISTEN	=>	GEWÄHRLEISTUNG

- (einen Versuch) eine(r) inhaltlichen Klassifizierung des Wortmaterials nach Sachwelten (EDV-Welt, EDV-Rechte-Welt, RechtsWelt, ...) für den hier vorgegebenen Bereich, der v.a. den Fachmann (Juristen, EDVler) einen Pfad (sachgerechtes Browning) durch die Datensammlung anbieten soll.

  Beispiel:

a)	Datenträger (EDV)	=>	Band(i),
		=>	Platte (j)
b)	Datum (RECHT)	=>	Personenbezogenes Datum
			=> geschütztes Datum
			=> freies Datum
		=>	Sachdatum

  Dies soll durch (eine Art von) facettierte(r) hierarchische(r) Klassifikation bewirkt werden, die dem Benutzer nicht notwendig bekannt sein muss. Daneben werden die üblichen Relationen (vgl. DIN 1463) etabliert.

Die Strukturen werden nun nicht nur beim Retrieval, sondern auch nach der Textanalyse angesprochen; die Daten befinden sich dazu in einer so genannten "Identifikations- und Zerlegungsdatei" (IuZ-Datei); der Zugriff geschieht über den "Lemmanamen" eines Textelements (also z.B. die Grundform eines Nomens, ggf. ergänzt um bedeutungsdifferenzierende Kennungen, wenn solche Differenzierungen schon durch die Analyse erzielt werden konnten.

Wird eine Mehrdeutigkeit ermittelt, die nicht reduziert werden konnte, so kann diese nun intellektuell (maschinengestützt - im Dialog -) vorgenommen werden; wird dies nicht ausgeführt bzw. ist eine Auflösung nicht möglich, so bleibt die Mehrdeutigkeit - entsprechend markiert - für die Dokument-Deskribierung erhalten.

Dem Benutzer sollen - abgesehen von den semantischen Verknüpfungen zunächst folgende "Deskriptortypen" zur Verfügung gestellt werden:

(1) "Normaldeskriptoren", d.h. solche (semantisch eindeutige oder vereindeutigte) Wörter oder Wortfolgen, die in der IuZ-Datei "lexikalisiert" sind (ausgeklammert sind u.a. Funktionswörter wie UND oder ALS; Homonyme sind durch Kennungen differenziert.

([1)] Technisch kann der Benutzer bei Mehrdeutigkeiten über ein Informationsdokument die Form des vereindeutigten Deskriptors erfragen.)

Beispiele:

ORGANISATION (ORDNUNG) [1)]
BUNDESRAT
DV-ANLAGE
MASSNAHMEN ERGREIFEN
PERSONENBEZOGENE DATEN
BUNDESDATENSCHUTZGESETZ
BDSG
OBERSTE BUNDESBEHÖRDE
AMT (BEHOERDE)
STELLE (INSTITUTION)

(2) "Teilwortdeskriptoren", d.h. Deskriptoren, die Ergebnis einer (semantisch sinnvollen) Zerlegung eines zusammengesetzten Deskriptors sind:

Beispiele:	aus:
DATENVERARBEITUNG SYSTEM	DATENVERARBEITUNGSSYSTEM
REGIERUNG BUND (i)	BUNDESREGIERUNG

(3) "Mehrdeutige Deskriptoren", d.h. solche, die während des Verarbeitungsprozesses (noch) nicht vereindeutigt werden konnten (will man sie vermeiden, muss ggf. intellektuell in den Analyseprozess eingegriffen werden) Beispiel: AMT, BAUER

(4) Pseudodeskriptoren", d.h. Wörter, die ein semantisch nicht differenziertes Wort (BAUER, AMT, ANLAGE) darstellen. Dies soll dem Benutzer erlauben, vorgenommene semantische Differenzierungen außer Acht zu lassen.

(5) Deskriptoren mit linguistisch ermittelten Verknüpfungen. Hier soll zumindest eine Auswahl aufgrund der Vermutung getroffen werden, dass solche Oberflächenstrukturen in

dem Fachtext besonders interessant sind, die häufig auch in "festen Wendungen" (FW) realisiert werden.

Nach dem Muster "Adjektiv-Attribut-Nomen" (FW: "juristische Person") werden alle attributiven und auch die erkannten prädikativen Adjektiv-Nomen-Verbindungen (GESCHÜTZT + DATEN; SEIN GESCHÜTZT + DATEN) verknüpft und retrievalfähig (zunächst nur auf Teilwortschatz beschränkt). Dies reduziert - nebenbei bemerkt - auch das Problem der Abgrenzung von entsprechenden festen Wendungen.

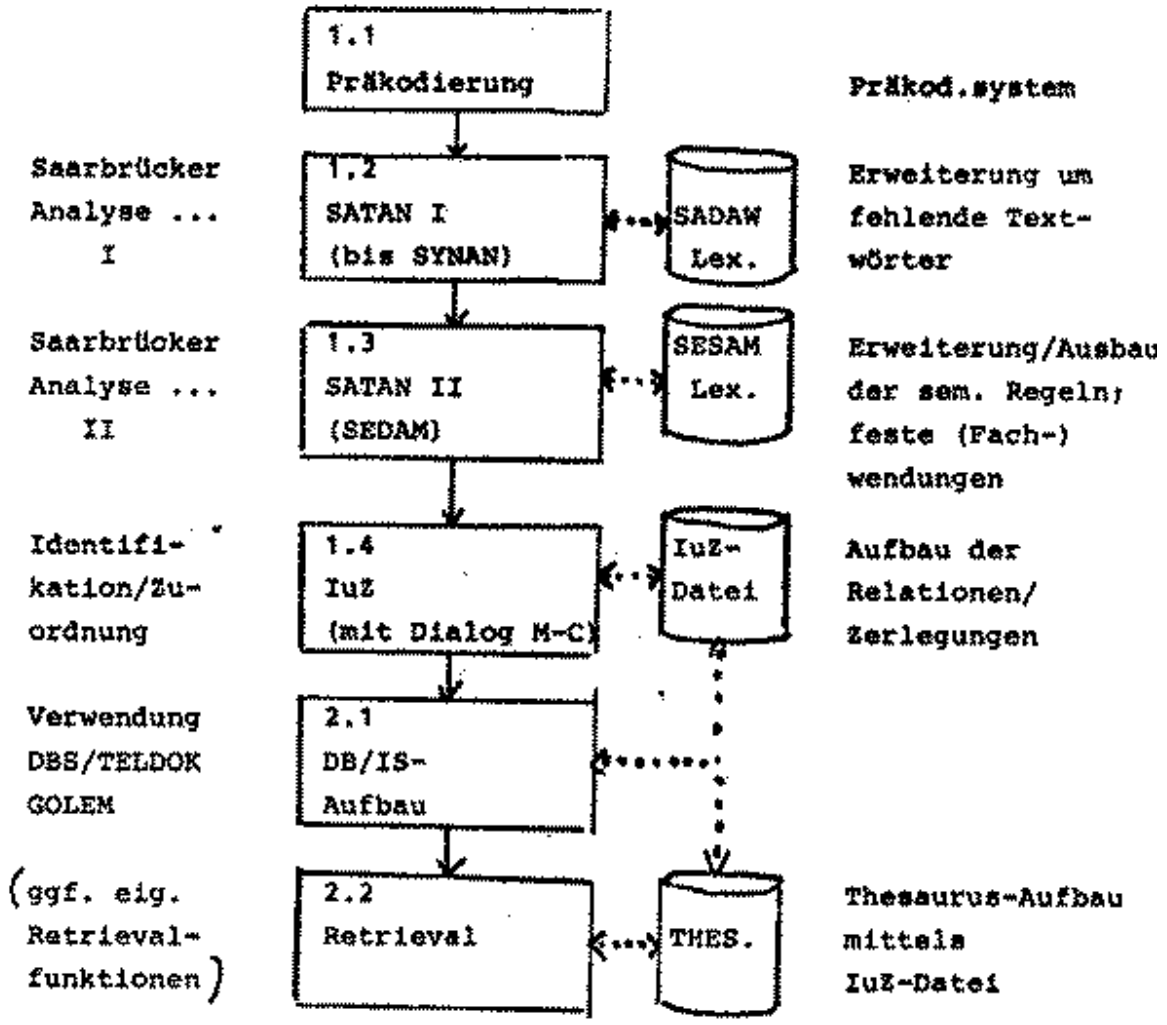
Als zweites soll die attributive Nomen-Nomen-Verknüpfung realisiert werden (s.o.), wobei ggf. noch Relatoren differenziert werden können:

ZUSTIMMUNG (GEN) BUNDESRAT
AUSFÜHRUNG (GEN) VORSCHRIFT
MASSNAHME (ZU) GEWÄHRLEISTUNG
GEWÄHRLEISTUNG (GEN) GESETZ
GESCHÜTZT (ATT) DATUM (INF)
(SEIN) ERFORDERLICH (ATT) MASSNAHME
AUSFÜHRUNG (GEN) VORSCHRIFT
ANLAGE (ZU) GESETZ
ZUSTIMMUNG (GEN) BUNDESRAT

Das Host-Retrieval-System wird unter den heute vorhandenen ausgewählt; vorgesehen sind eine TELDOK- und eine GOLEM-Version; daneben ist - falls sich dies im Rahmen der Projektfinanzierung noch ermöglichen lässt - eine Retrieval-Version auf DBS-Grundlage geplant, die den vorgesehenen speziellen Funktionen besser gerecht wird.

2.7 Das Projekt soll 1979 abgeschlossen werden; es werden durchschnittlich 4 hauptamtliche Mitarbeiter beschäftigt, so dass die Modellversion mit einem Aufwand von ca. 10-12 Mannjahren realisiert wird. Dies wäre ohne eine intensive Ausnutzung (und Unterstützung) der Saarbrücker Analyse und vorhandener Retrieval-Systeme nicht möglich.

JUDO - Bausteine



3 ENZY: Automatische Verarbeitung enzyklopädischer Lexika

ENZY stellt ein relativ kleines Forschungsprojekt dar, das in Kooperation mit Siemens (Finanzgeber) und dem Bibliographischen Institut (Datenlieferant) durchgeführt. wird.

3.1 In einer ersten Phase wurden Stichwörter aus "Meyers Enzyklopädisches Lexikon" (das ab Band 9 über Lichtsatz produziert wird und auf Lochstreifen vorliegt) daraufhin untersucht, ob die Artikel maschinell in kleinere Einheiten zerlegt und ggf. Teilmengen daraus automatisch strukturiert werden könnten.

Am Beispiel von Biographien (Kopfteil) konnte dies ziemlich gut durchgeführt werden, da diese genügend formale Kennzeichen (wie "Geboren-Stern") aufweisen.

Besondere Probleme machte die Mehrdeutigkeit des Punkts (Dezimal-, Satzende-, Abkürzungspunkt ...); es zeigte sich insgesamt, dass eine Differenzierung bereits bei der Erfassung wohl gegenwärtig der einzige Weg ist, datenbankorientierte Verarbeitungen (etwa im Hinblick auf verbesserte Update-Möglichkeiten oder gezielte Datenauswahl für Speziallexika) zu ermöglichen

- übrigens ein "Aufwand", dem sich die Redaktionen - da vordergründig kostenverursachend - nicht gerne unterziehen.

3.2 In der zweiten Projektphase wurde daher zunächst versucht, das "Vertrauen" der Redaktion durch ein "Programm zur Pflege und Ermittlung von Verweisen" (auf ein Stichwort aus einem anderen Stichwort) zu gewinnen. Dies soll die Redaktion bei ihrer Arbeit unmittelbar unterstützen (und z.B. Zirkelverweise vermeiden helfen). Ein besonderes Problem ist dabei die Erkennung mehrwortiger Verweise; dies gelingt aufgrund formaler Kriterien jedoch überraschend zuverlässig, wenn auch die Verweisinformation dann nicht "lemmatisiert" ist.

3.3 Ein Hauptproblem ist - nebenbei bemerkt - die Bewältigung des Zeichenvorrats; "Meyers Enzyklopädisches Lexikon" kennt über 2000 Zeichendarstellungen, wenn man die Schriftarten "halbfett" und "kursiv" mit einbezieht: es kommen phonetische, griechische und mathematische Zeichen (neben den für die Satzsteuerung nötigen) vor.

3.4 Mittelfristig soll das Ziel einer systematisch(er)en Erfassung der Enzyklopädischen Daten weiter verfolgt werden; Grundlage werden "Checklisten," und. Formate sein, die objekttyp-spezifisch und fachgebietsorientiert sind. Wenn zugleich einige formale Regeln (z.B. weniger Abkürzungen, Vereindeutigung der Punkte) eingehalten werden, sind sicherlich auch maschinelle Sprachanalyseverfahren einsetzbar; allerdings sind die zu verarbeitenden Datenmengen derart umfangreich, dass eine sachliche Beschränkung erforderlich sein wird.

3.5 Das Projekt wird Mitte 1975 gefördert und erforderte bislang einen Finanzierungsaufwand von einem Mannjahr.

Literatur:

Krause, 3. 1977: JUDO: Ein Verfahren zur automatischen Indexierung juristischer Texte, 7. Jahrestagung der Gesellschaft für Informatik. JUDO-V-02.

Krause, J. 1978: Textwörterbücher, maschinelle Sprachanalyse und Dokumentation (LDV-Aktivitäten in Regensburg, 2. Teil).

Zimmermann, H. 1976: Linguistische Informationswissenschaft. In: Nachr. Dok . 27 S. 133-135.

Zimmermann, H. 1977 a: Juristische Dokumentanalyse (JUDO). Vorhabensbeschreibung. Arbeitspapier JUDO-A-01.

Zimmermann, H. 1977 b: Probleme der automatischen Indexierung von Fachtexten am Beispiel juristischer Dokumente, Frühjahrstagung des FB "Medizinische Informatik". JUDO-V-01. Jetzt auch in: Medizinische Informatik und Statistik 4, Klartextverarbeitung, ed. F. Wingert, Berlin-Heidelberg-New York 1978, S. 112 - 121.